

# CONVERGENCE OF STOCHASTIC PROXIMAL GRADIENT ALGORITHM

LORENZO ROSASCO <sup>†\*</sup>, SILVIA VILLA <sup>†</sup>, AND BÀNG CÔNG VŨ <sup>†</sup>

**Abstract.** We prove novel convergence results for a stochastic proximal gradient algorithm suitable for solving a large class of convex optimization problems, where a convex objective function is given by the sum of a smooth and a non-smooth component. We derive convergence rates in expectation in the strongly convex case, as well as almost sure convergence results under weaker assumptions.

**Key words.** Proximal Methods, Forward-backward splitting algorithm, Stochastic optimization, Online Learning Algorithms.

**1. Introduction.** First order methods have been recently widely applied to solve convex optimization problems in a variety of areas including machine learning and signal processing. In particular, proximal gradient algorithms (a.k.a. forward-backward splitting algorithms) and their accelerated variants have received considerable attention ([4, 16, 31, 5] and references therein). These algorithms are easy to implement and suitable for solving high dimensional problems thanks to the low memory requirement of each iteration. Moreover, they are particularly suitable for composite optimization, where a convex objective function is the sum of a smooth and a non-smooth component. This class of optimization problems arises naturally in regularization schemes where the smooth term is a data fitting term and the non-smooth term a regularizer, see for example [14, 28]. Interestingly, proximal splitting algorithms separate the contribution of the components of the objective function at each iteration: the proximal operator defined by the non smooth term is applied to a gradient descent step for the smooth term. Recent studies have considered the case where the proximal operator is known only up-to an error [36, 41], while the gradient of the smooth term is usually assumed to be known exactly. In fact, it is often the case that only a stochastic estimate of the gradient is available and it is therefore interesting to develop stochastic versions of proximal splitting methods.

The study of classical stochastic approximation methods originates in the work of [35], and assumes the objective function to be smooth and strongly convex; the related literature is large (see e.g. [7, 29, 18] and references therein). An improvement of the original stochastic approximation method, which can be extended to nonsmooth convex minimization, is proposed by [30] and [32]. Their method relies on the averaging of the trajectories and allow for larger step-sizes. In [25], variants of the mirror descent stochastic approximation algorithm [22] have been proposed to solve stochastic composite optimization problems. In particular, [25] proposes an accelerated method and derive a rate of convergence for the objective function values which is optimal both with respect to the smooth component and the non-smooth term. Similar accelerated proximal gradient algorithms have been also studied in the machine learning community, see [24, 42], and also [11, 37, 38, 43].

Our stochastic proximal gradient algorithm is related to the FOBOS algorithm presented in [19]. Here, we allow for an additional relaxation step, which in practice can speed up convergence, and derive novel results. Differently from most previous

\* DIBRIS, Università di Genova, Via Dodecaneso, 35, 16146, Genova, Italy, (lrosasco@mit.edu)

† LCSL, Istituto Italiano di Tecnologia and Massachusetts Institute of Technology, Bldg. 46-5155, 77 Massachusetts Avenue, Cambridge, MA 02139, USA, (Silvia.Villa@iit.it, Cong.Bang@iit.it)

works, we focus on almost sure convergence and convergence in expectation of the iterates, rather than the function values, and consider an infinite dimensional setting. Our convergence results do not require averaging, and allow the choice of step-sizes of the form  $n^{-\theta}$  with  $\theta \in ]0, 1]$ . This is relevant, since, as pointed out in [33, 39], averaging can have a negative impact in the strongly convex case, and, most importantly, if sparsity based regularization is considered, it typically prevents obtaining sparse solutions. Another contribution of our approach is the possibility of having unbounded stochastic estimates of the gradients. More in details, the analysis in the paper is divided in two parts. In the first, we study convergence in expectation in the strongly convex case, generalizing the results in [2, Section 3] to the nonsmooth case. We provide a non-asymptotic analysis of stochastic proximal gradient descent where the bounds depend explicitly on the parameters of the problem. The optimal  $O(1/n)$  convergence rate is achieved. In the second part, we establish almost sure convergence. The study of almost sure convergence has a long history, see e.g. [34, 23, 7, 12] and references therein. Other recent results on almost sure convergence of projected stochastic gradient algorithm can be found in [6, 27], under rather technical assumptions. Our results are generalizations to the composite case of the analysis of the stochastic projected subgradient algorithm in an infinite dimensional Hilbert space in [3].

The paper is organized as follows. In Section 2 we introduce composite optimization and the stochastic proximal gradient algorithm, along with some relevant special cases. In Section 3, we study convergence in expectation and almost surely that we prove in Section 4. Auxiliary results are found in Appendix A, while Appendix B considers the particular case of minimization over orthonormal bases [14].

Note that this paper was submitted for publication on February 7th, 2014 and is currently under review. A related paper has since appeared on the arxiv [1].

**Notation.** Throughout,  $(\mathbf{E}, \mathcal{A}, \mathbf{P})$  is a probability space,  $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$ , and  $\mathcal{H}$  a real separable Hilbert space. We use the notation  $\langle \cdot, \cdot \rangle$  and  $\| \cdot \|$  for the scalar product and the associated norm in  $\mathcal{H}$ . The symbols  $\rightharpoonup$  and  $\rightarrow$  denote, respectively, weak and strong convergence. The class of lower semicontinuous convex functions  $f: \mathcal{H} \rightarrow ]-\infty, +\infty]$  such that  $\text{dom } f = \{x \in \mathcal{H} \mid f(x) < +\infty\} \neq \emptyset$ , is denoted by  $\Gamma_0(\mathcal{H})$ . The proximity operator of  $f \in \Gamma_0(\mathcal{H})$  is denoted by  $\text{prox}_f$  (see (A.2)). We denote by  $\sigma(X)$  the  $\sigma$ -field generated by a random variable  $X$ . The expectation of a random variable  $X$  is denoted by  $\mathbb{E}[X]$ . The conditional expectation of  $X$  given a  $\sigma$ -field  $\mathcal{F} \subset \mathcal{A}$  is denoted by  $\mathbb{E}[X|\mathcal{F}]$ . The conditional expectation of  $X$  given  $Y$  is denoted by  $\mathbb{E}[X|Y]$ . The shorthand notation 'a.s.' stands for 'almost sure'. Finally, we define the family of functions

$$\varphi_\beta: ]0, +\infty[ \rightarrow \mathbb{R}: t \mapsto \begin{cases} \beta^{-1}(t^\beta - 1) & \text{if } \beta \neq 0; \\ \log t & \text{if } \beta = 0. \end{cases}$$

**2. Problem setting and examples.** In this section, we introduce the composite convex optimization problem, the stochastic proximal method we study, and discuss some special cases of the framework we consider.

**2.1. Problem.** Composite optimization problems are defined as the the problem of minimizing the sum of a smooth convex function and a nonsmooth convex function which is proximable, that is the proximity operator (A.2) can be computed.

**PROBLEM 2.1.** *Let  $R: \mathcal{H} \rightarrow ]-\infty, +\infty]$  be proper and lower semicontinuous, and let  $L: \mathcal{H} \rightarrow \mathbb{R}$  be convex and differentiable, with a  $\beta^{-1}$ -Lipschitz continuous gradient,*

$\beta \in ]0, \infty[$ . The problem is to

$$\underset{w \in \mathcal{H}}{\text{minimize}} T(w) = L(w) + R(w), \quad (2.1)$$

under the assumption that the set of solutions to (2.1) is non-empty. As we mentioned in the introduction, this problem is classical and has been extensively studied in convex optimization. In this paper, we study the following stochastic proximal gradient algorithm (SPGA).

ALGORITHM 2.2 (SPGA). Let  $(\gamma_n)_{n \in \mathbb{N}^*}$  be a strictly positive sequence and  $(\lambda_n)_{n \in \mathbb{N}^*}$  be a sequence in  $[0, 1]$ , let  $(G_n)_{n \in \mathbb{N}^*}$  be a  $\mathcal{H}$ -valued random process. Fix  $w_1 \in \mathcal{H}$  and set

$$(\forall n \in \mathbb{N}^*) \quad \begin{cases} z_n = w_n - \gamma_n G_n \\ y_n = \text{prox}_{\gamma_n R} z_n \\ w_{n+1} = (1 - \lambda_n) w_n + \lambda_n y_n. \end{cases} \quad (2.2)$$

Algorithm 2.2 is a stochastic version of the proximal forward-backward splitting [4], where we replace the exact gradient by a stochastic estimate. More specifically, when  $(\forall n \in \mathbb{N}) G_n = \nabla L(w_n)$ , our algorithm reduces to the one in [16]. A stochastic proximal forward-backward splitting (FOBOS) was firstly proposed in [19] for minimizing the sum of two functions where one of them is proximable, and the other is convex and subdifferentiable. Algorithm 2.2 generalizes the FOBOS algorithm, by including a relaxation step (while assuming the first component in (2.1) to be smooth). As it is the standard, to ensure convergence of the proposed algorithm, we need additional conditions on the stochastic gradient  $(G_n)_{n \in \mathbb{N}^*}$  as well as on the step-sizes  $(\gamma_n)_{n \in \mathbb{N}^*}$ .

CONDITION 2.3. The following conditions will be considered for the filtration  $(\mathcal{F}_n)_{n \in \mathbb{N}^*}$  with  $\mathcal{F}_n = \sigma(w_1, \dots, w_n)$ .

(A1) For every  $n \in \mathbb{N}^*$ ,  $\mathbb{E}[G_n | \mathcal{F}_n] = \nabla L(w_n)$ .

(A2) For every  $n \in \mathbb{N}^*$ ,  $\mathbb{E}[\|G_n - \nabla L(w_n)\|^2 | \mathcal{F}_n] \leq \sigma^2(1 + \alpha_n \|\nabla L(w_n)\|^2)$ , for some  $\alpha_n \in ]0, +\infty[$ .

(A3) For some  $\epsilon \in ]0, +\infty[$ , it holds  $(\forall n \in \mathbb{N}^*) 1 - \frac{\gamma_n(1+2\sigma^2\alpha_n)}{\beta} \geq \epsilon > 0$

(A4) Summability: For any solution  $\bar{w}$  of the problem (2.1),

$$\sum_{n \in \mathbb{N}^*} \lambda_n \gamma_n = +\infty \quad \text{and} \quad \sum_{n \in \mathbb{N}^*} \chi_n^2 < +\infty, \quad \text{where} \quad \chi_n^2 = \lambda_n \gamma_n^2 (1 + 2\alpha_n \|\nabla L(\bar{w})\|^2). \quad (2.3)$$

Condition (A1) means that, at each iteration, we have an unbiased estimate of the gradient of the smooth term. Condition (A2) has been considered in [3]. Assumption (A2) is weaker than the ones commonly used in the analysis of stochastic (sub)gradient algorithms, namely uniform boundedness of  $\mathbb{E}[\|G_n\|^2 | \mathcal{F}_n]$  (see [29]) or even boundedness of  $\|G_n\|^2$  (see [19]). We note that this last requirement on the entire space is not compatible with the assumption of strong convexity, because the gradient is necessarily not uniformly bounded, therefore the use of the more general condition (A2) is needed. Another possibility is to require the boundedness assumption only on a ball containing the unique solution. The ball can be shown to contain the iterates for a suitable radius depending on the strong convexity constant (see [19, Lemma 9 and Corollary 10]).

Conditions such as (A3) and (A4) are, respectively, widely used in the deterministic setting and in stochastic optimization. We also note that when  $(\lambda_n)_{n \in \mathbb{N}^*}$  is bounded away from zero, and  $(\alpha_n)_{n \in \mathbb{N}^*}$  is bounded, (A4) implies (A3) for  $n$  large enough. The

condition  $\sum_{n \in \mathbb{N}^*} \lambda_n^2 < +\infty$  in Assumption (A4) is satisfied if  $(\lambda_n \gamma_n^2 (1 + 2\alpha_n))_{n \in \mathbb{N}^*}$  is summable. Moreover, if  $G = 0$ , it reduces to  $\sum_{n \in \mathbb{N}^*} \lambda_n \gamma_n^2 < +\infty$ , since in this case  $\nabla L(\bar{w}) = 0$  for every solution  $\bar{w}$ . We note that the assumptions on the step-size are different from those usually made in the deterministic setting. In our case, the step-size is required to converge to zero, while it is bounded away from zero in [16].

**2.2. Special cases.** Problem 2.1 covers a wide class of deterministic as well as stochastic convex optimization problems, see e.g. [14, 29] and references therein. The simplest case is when  $R$  is identically equal to 0, so that Problem 2.1 reduces to the classic problem of finding a minimizer of a convex differentiable function from unbiased estimates of its gradients. In the case when  $R$  is the indicator function of a nonempty, convex, closed set  $C$ , i.e.

$$R(w) = \iota_C(w) = \begin{cases} 0 & w \in C, \\ +\infty & w \notin C, \end{cases}$$

then problem (2.1) reduces to a constrained minimization problem of the form

$$\underset{w \in C}{\text{minimize}} \quad L(w),$$

which is well studied in the literature, as mentioned in the introduction. Below, we discuss in more detail some special cases of interest.

**EXAMPLE 1. (Minimization of an Expectation).** Let  $\xi$  be a random vector with probability distribution  $P$  supported on  $\mathbf{E}$  and  $F: \mathcal{H} \times \mathbf{E} \rightarrow \mathbb{R}$ . A standard setting for studying stochastic gradient descent methods is when

$$L(w) = \mathbb{E}[F(w, \xi)] = \int_{\mathbf{E}} F(w, \xi) dP(\xi),$$

under the assumption that  $(\forall \xi \in \mathbf{E}) F(\cdot, \xi)$  is a convex differentiable function with Lipschitz continuous gradient [29]. Let  $(\xi_n)_{n \in \mathbb{N}^*}$  be independent copies of the random vector  $\xi$  whose probability distribution is supported on a probability space  $\mathbf{E}$ . Assume that there is an oracle that, for each  $(w, \xi) \in \mathcal{H} \times \mathbf{E}$ , returns a vector  $G(w, \xi)$  such that  $\nabla L(w) = \mathbb{E}[G(w, \xi)]$ . By setting  $(\forall n \in \mathbb{N}^*) G_n = G(w_n, \xi_n)$  and  $\mathcal{F}_n = \sigma(\xi_1, \dots, \xi_n)$ , then (A2) holds. This latter assertion follows from standard properties of conditional expectation, see e.g. [20, Example 5.1.5].

**EXAMPLE 2. (Minimization of a Sum of Functions)** Let  $R \in \Gamma_0(\mathcal{H})$ , let  $m$  be a strictly positive integer. For every  $i \in \{1, \dots, m\}$ , let  $L_i$  be convex and differentiable, such that  $\sum_{i=1}^m L_i$  has a  $\beta^{-1}$ -Lipschitz continuous gradient, for some  $\beta \in ]0, +\infty[$ . The problem is to

$$\underset{w \in \mathcal{H}}{\text{minimize}} \quad \frac{1}{m} \sum_{i=1}^m L_i(w) + R(w).$$

This problem is a special case of Problem 2.1 with  $L = \sum_{i=1}^m L_i$ , and is especially of interest when  $m$  is very large and we know the exact gradient of each component  $L_i$ . The stochastic estimate of the gradient of  $L$  is then defined as

$$(\forall n \in \mathbb{N}) \quad G_n = \nabla L_{i(n)}(w_n), \quad (2.4)$$

where  $i$  is a discrete random variable from  $\mathbb{N}^*$  to  $\{1, \dots, m\}$ , see [8, 9]. Clearly (A1) holds. If assumption (A2) is satisfied, then SPGA can be applied.

Finally, in the next section, we discuss how the above setting specializes to the context of machine learning.

**2.3. Application to Machine Learning.** Consider two measurable sets  $\mathcal{X}$  and  $\mathcal{Y}$  and assume there is a probability measure  $\rho$  on  $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ . The distribution  $\rho$  is fixed but known only through a training set  $\mathbf{z} = (x_i, y_i)_{1 \leq i \leq m} \in \mathcal{Z}^m$  of samples i.i.d with respect to  $\rho$ . Consider a loss function  $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, +\infty[$  and a hypothesis space  $\mathcal{H}$  of functions from  $\mathcal{X}$  to  $\mathcal{Y}$ , e.g. a reproducing kernel Hilbert space. A problem relevant in this context is (regularized) empirical risk minimization,

$$\underset{w \in \mathcal{H}}{\text{minimize}} \quad \frac{1}{m} \sum_{i=1}^m \ell(y_i, w(x_i)) + R(w), \quad (2.5)$$

The above problem can be seen as an approximation of the problem,

$$\underset{w \in \mathcal{H}}{\text{minimize}} \quad \int_{\mathcal{X} \times \mathcal{Y}} \ell(y, w(x)) d\rho + R(w). \quad (2.6)$$

The analysis, in this paper, can be adapted to the machine learning setting in two different ways. The first, following Example 1, is to apply the SPGA to directly solve the regularized *expected* risk minimization problem (2.6). The second, following Example 2, is to apply the SPGA to solve the regularized *empirical* risk minimization problem (2.5).

In either one of the above two problems, the first term is differentiable if the loss functions is differentiable with respect to its second argument, examples being the squared or the logistic loss. For these latter loss functions, and more generally for loss functions which are twice differentiable in their second argument, it is easy to see that the Lipschitz continuity of the gradient is satisfied under mild boundedness assumptions. The second term  $R$  can be seen as a regularizer/penalty encoding some prior information about the learning problem. Examples of convex, non-differentiable penalties include sparsity inducing penalties such as the  $\ell_1$  norm, as well as more complex structured sparsity penalties [28]. Stronger convexity properties can be obtained considering an *elastic net penalty* [44, 17], that is adding a small strongly convex term to the sparsity inducing penalty. Clearly, the latter term would not be necessary if the error term in Problem 2.6 (or in (2.5)) is strongly convex. However, this latter requirement depends on the probability distribution  $\rho$  and is typically not satisfied when considering high (possibly infinite) dimensional settings.

**3. Main results and discussion.** In this section, we state and discuss the main results of the paper. We derive convergence rates of the proximal forward-backward splitting (with relaxation) for stochastic minimization. The section is divided in two parts. In the first one, Section 3.1, we focus on convergence in expectation. In the second one, Section 3.2, we study almost sure convergence of the sequence of iterates. In both cases, additional convexity conditions on the objective function are required to derive convergence results. The proofs are deferred to Section 4.

**3.1. Convergence in Expectation of SPGA.** In this section, we denote by  $\bar{w}$  the solution of Problem 2.1 and provide an explicit non-asymptotic bound on  $\mathbb{E}[\|w_n - \bar{w}\|^2]$ . This result generalizes to the nonsmooth case the bound obtained in [2, Theorem 1] for stochastic gradient descent. The following assumption is considered throughout this section.

**ASSUMPTION 3.1.** *The function  $L$  is  $\mu$ -strongly convex and  $R$  is  $\nu$ -strongly convex, for some  $\mu \in [0, +\infty[$  and  $\nu \in [0, +\infty[$ , with  $\mu + \nu > 0$ .* Note that, we do not assume both  $L$  and  $R$  to be strongly convex, indeed the constants  $\mu$  and  $\nu$  can be zero,

but require that only one of the two is. This implies that Problem 2.1 has a unique solution, denoted by  $\bar{w}$ .

**THEOREM 3.2.** *Assume that conditions (A1), (A2), and (A3) and Assumption 3.1 are satisfied. Let  $\inf_{n \in \mathbb{N}^*} \lambda_n \geq \bar{\lambda} > 0$ ,  $\sup_{n \in \mathbb{N}^*} \alpha_n \leq \bar{\alpha} < +\infty$  and  $\gamma_n = c_1 n^{-\theta}$  for some  $\theta \in ]0, 1]$  and for some  $c_1 \in ]0, +\infty[$ . Set  $t = 1 - 2^{\theta-1} \geq 0$ ,  $c = 2c_1 \underline{\lambda}(\nu + \mu\varepsilon)/(1 + \nu)^2$ ,  $\tau = 2\sigma^2 c_1^2(1 + \bar{\alpha}\|\nabla L(\bar{w})\|)/c^2$  and let  $n_0$  be the smallest integer such that  $(\forall n \geq n_0 > 1) \max\{c, c_1\}n^{-\theta} \leq 1$ . Then, by setting*

$$(\forall n \in \mathbb{N}^*) \quad s_n = \mathbb{E} [\|w_n - \bar{w}\|^2],$$

we have, for every  $n \geq 2n_0$ ,

$$s_{n+1} \leq \begin{cases} \left( \tau c^2 \varphi_{1-2\theta}(n) + s_{n_0} \exp\left(\frac{cn_0^{1-\theta}}{1-\theta}\right) \right) \exp\left(\frac{-ct(n+1)^{1-\theta}}{1-\theta}\right) + \frac{\tau 2^\theta c}{(n-2)^\theta} & \text{if } \theta \in ]0, 1[, \\ s_{n_0} \left(\frac{n_0}{n+1}\right)^c + \frac{\tau c^2}{(n+1)^c} \left(1 + \frac{1}{n_0}\right)^c \varphi_{c-1}(n) & \text{if } \theta = 1. \end{cases} \quad (3.1)$$

In Theorem 3.2, the dependence on the strong convexity constants is hidden in the constant  $c$ . Taking into account (3.1), we can write more explicitly the asymptotic behavior of the quantity  $s_n$ .

**COROLLARY 3.3.** *Under the same assumptions and with the same notation as in Theorem 3.2, the following holds*

$$\mathbb{E}[\|w_n - \bar{w}\|^2] = \begin{cases} O(n^{-\theta}) & \text{if } \theta \in ]0, 1[, \\ O(n^{-c}) + O(n^{-1}) & \text{if } \theta = 1. \end{cases} \quad (3.2)$$

In particular, if  $\theta = 1$  and  $c_1 \geq (1 + \nu)^2/(2\underline{\lambda}(\nu + \mu\varepsilon))$  then  $c \geq 1$ , and  $\mathbb{E}[\|w_n - \bar{w}\|^2] = O(n^{-1})$ .

As can be seen from Corollary 3.3, the fastest asymptotic rate corresponds to  $\theta = 1$  and it is the same obtained in the smooth case in [2, Theorem 2]. Note that this rate does not only depend on the asymptotic behavior of the step-size, but also on the constant  $c$ , which in turns depends on  $c_1$ . As pointed out in [29], see also in [2], this choice is critical, because too small choices of  $c_1$  affect the convergence rates, and too big choices influence significantly the value of the constants in the first term of (3.1). In particular, the choice is determined by available estimates on the strong convexity constants.

Theorem 3.2 is the extension to the nonsmooth case of [2, Theorem 1], in particular, when  $R = 0$ , we obtain the same convergence rate. Note however that the assumptions on the stochastic approximations of the gradient of the smooth part are slightly different. In particular, we replace the boundedness condition at the solution and the Lipschitz continuity assumption on  $G_n$  with assumption (A2).

We briefly contrast our results with those in [19]. This latter paper considers convergence of the average of the iterates with respect to the function values and assume  $\mathcal{H}$  to be finite dimensional. Also uniform boundedness of the iterations and the subdifferentials are required. Our convergence results consider convergence of the iterates (with no averaging) and hold in an infinite dimensional setting, without boundedness assumptions. The non asymptotic rate  $O(n^{-1})$  which we obtain for the iterates improves the  $O((\log n)/n)$  rate derived from [19, Corollary 10] for the average of the iterates. However, it should be noted that convergence of the objective values

is studied in [19] also for the non strongly convex case.

Theorem 3.2 is also comparable with deterministic stochastic proximal forward-backward algorithm with errors [16]. On the one hand, we allow the errors to grow according to assumption (A2), while in the deterministic case the errors in the computation of the gradient should decrease to zero sufficiently fast. On the other hand, we require asymptotically vanishing step-sizes, while, in the deterministic case, the step-size is bounded from below.

**3.2. Almost sure convergence of SPGA.** In this section, we focus on almost sure convergence of SPGA. This kind of convergence of the iterates is the one traditionally studied in the stochastic optimization literature. Depending on the convexity properties of the function  $L$ , we get two different convergence properties. The first theorem requires uniform convexity of  $L$  at the solution.

**THEOREM 3.4.** *Suppose that the conditions (A1), (A2), (A3), and (A4) are satisfied. Let  $(w_n)_{n \in \mathbb{N}^*}$  be a sequence generated by Algorithm 2.2 and assume that  $L$  is uniformly convex at  $\bar{w}$ . Then  $w_n \rightarrow \bar{w}$  a.s. If we relax the strong convexity assumption, we can still prove weak convergence of a subsequence in the strictly convex case, provided an additional regularity assumption holds.*

**THEOREM 3.5.** *Suppose that the conditions (A1), (A2), (A3), and (A4) are satisfied. Let  $(w_n)_{n \in \mathbb{N}^*}$  be a sequence generated by Algorithm 2.2. Assume that  $L$  is strictly convex, so that Problem 2.1 has a unique solution  $\bar{w}$ . If  $\nabla L$  is weakly continuous, then there exists a subsequence  $(w_{t_n})_{n \in \mathbb{N}^*}$  such that  $w_{t_n} \rightharpoonup \bar{w}$  a.s.*

With respect to the previous section, here we make the additional assumption (A4) on the summability of the sequence of step-sizes multiplied by the relaxation parameters. For stochastic gradient descent without relaxation, i.e,  $R = 0$  and  $(\forall n \in \mathbb{N}^*) \lambda_n = 1$ , assumption (A4) coincides with the classical step-size condition  $\sum_{n \in \mathbb{N}^*} \gamma_n = +\infty$  and  $\sum_{n \in \mathbb{N}^*} \gamma_n^2 < +\infty$  which guarantees a sufficient but not too fast decrease of the step-size (see e.g. [10]). Assumption (A2) has been considered in the context of stochastic gradient descent in [10]. Note that under such a condition, the variance of the stochastic approximation is allowed to grow with  $\|\nabla L(w_n)\|$  and therefore can be unbounded.

As we mentioned in the introduction, the study of almost sure convergence has a long history. However, most papers consider only the finite dimensional setting. An analysis of a stochastic projected subgradient algorithm in an infinite dimensional Hilbert space can be found in [3]. Theorem 3.4 can be seen as an extension of [3, Theorem 3.1], where the case where  $R$  is an indicator function is considered. As in [3], our approach is based on probabilistic quasi martingale techniques [26].

**REMARK 1.** *If  $L$  is assumed to be only strictly convex and its gradient is not weakly continuous, (the previous results do not ensure weak convergence of any subsequence of  $(w_n)_{n \in \mathbb{N}^*}$ . However, if one can show that the sequence of function values  $(T(w_n))_{n \in \mathbb{N}^*}$  converges to the optimal value, then  $w_n \rightarrow \bar{w}$  a.s. This happens (see [3]) when  $R = \iota_V$  for some closed subspace  $V$  of  $\mathcal{H}$ , or when  $R = \iota_C$  for some non-empty closed convex  $C$  of  $\mathcal{H}$ , and there exists a bounded function  $h: \mathbb{R} \rightarrow \mathbb{R}$  such that  $(\forall n \in \mathbb{N}) \mathbb{E}[\|G_n - \nabla L(w_n)\| | \mathcal{F}_n] \leq h(\|\nabla L(w_n)\|)$ . The proof of Remark 1 can be found in the appendix.*

**4. Proof of the Main Results.** We start by stating and proving some auxiliary results that will be useful in the proof of Theorems 3.2 and 3.4.

**PROPOSITION 4.1.** *Suppose that (A1), (A2), (A3), and (A4) are satisfied. Let  $(w_n)_{n \in \mathbb{N}^*}$  be a sequence generated by Algorithm 2.2. Then, for any solution  $\bar{w}$  of the problem (2.1), the following hold.*

- (i) *The sequence  $(\mathbb{E}[\|w_n - \bar{w}\|^2])_{n \in \mathbb{N}^*}$  converges to a finite value.*

(ii) The sequence  $(\|w_n - \bar{w}\|^2)_{n \in \mathbb{N}^*}$  converges a.s to some integrable random variable  $\zeta_{\bar{w}}$ .

(iii)  $\sum_{n \in \mathbb{N}^*} \lambda_n \gamma_n \mathbb{E}[\langle w_n - \bar{w}, \nabla L(w_n) - \nabla L(\bar{w}) \rangle] < +\infty$ . Consequently,

$$\lim_{n \rightarrow \infty} \mathbb{E}[\langle w_n - \bar{w}, \nabla L(w_n) - \nabla L(\bar{w}) \rangle] = 0 \quad \text{and} \quad \lim_{n \rightarrow \infty} \mathbb{E}[\|\nabla L(w_n) - \nabla L(\bar{w})\|^2] = 0.$$

(iv)  $\sum_{n \in \mathbb{N}^*} \lambda_n \mathbb{E}[\|w_n - y_n - \gamma_n(G_n - \nabla L(\bar{w}))\|^2] < +\infty$  and  $\sum_{n \in \mathbb{N}^*} \lambda_n \mathbb{E}[\|w_n - y_n\|^2] < +\infty$ .

*Proof.* Since  $\bar{w}$  is a solution of the problem (2.1), we have

$$(\forall n \in \mathbb{N}^*) \quad \bar{w} = \text{prox}_{\gamma_n R}(\bar{w} - \gamma_n \nabla L(\bar{w})). \quad (4.1)$$

Let us set

$$(\forall n \in \mathbb{N}^*) \quad y_n = \text{prox}_{\gamma_n R}(w_n - \gamma_n G_n) \quad \text{and} \quad u_n = w_n - y_n - \gamma_n(G_n - \nabla L(\bar{w})). \quad (4.2)$$

Then, it follows from the convexity of  $\|\cdot\|^2$  that

$$(\forall n \in \mathbb{N}^*) \quad \|w_{n+1} - \bar{w}\|^2 \leq (1 - \lambda_n)\|w_n - \bar{w}\|^2 + \lambda_n\|y_n - \bar{w}\|^2. \quad (4.3)$$

Since  $\text{prox}_{\gamma_n R}$  is firmly non-expansive by Lemma A.4, we have

$$\begin{aligned} (\forall n \in \mathbb{N}^*) \quad \|y_n - \bar{w}\|^2 &\leq \|(w_n - \bar{w}) - \gamma_n(G_n - \nabla L(\bar{w}))\|^2 - \|u_n\|^2 \\ &= \|w_n - \bar{w}\|^2 - 2\gamma_n \langle w_n - \bar{w}, G_n - \nabla L(\bar{w}) \rangle + \gamma_n^2 \|G_n - \nabla L(\bar{w})\|^2 - \|u_n\|^2. \end{aligned} \quad (4.4)$$

Note that, for every  $n \in \mathbb{N}^*$ , we have that  $w_n$  and  $G_n$  are measurable with respect to  $\mathcal{A}$  since they are  $\mathcal{F}_n$  measurable and by definition  $\mathcal{F}_n \subset \mathcal{A}$ . The same holds for  $y_n$ , for it is the difference of two measurable functions. We next show by induction that  $(\forall n \in \mathbb{N}^*)$   $w_n$  is integrable. First,  $w_1$  is integrable since it is constant by assumption. Then, assume by inductive hypothesis that  $w_n$  is integrable. Then so is  $y_n$ , for  $G_n$  is integrable by assumption. Moreover,  $\mathbb{E}[\|w_{n+1}\|] \leq (1 - \lambda_n)\mathbb{E}[\|y_n\|] + \lambda_n \mathbb{E}[\|w_n\|] \leq \mathbb{E}[\|w_n\|] + \|\text{prox}_{\gamma_n R} 0\|$ , because  $\text{prox}_{\gamma_n R}$  is nonexpansive. This implies that  $\mathbb{E}[\langle w_n - \bar{w}, G_n - \nabla L(\bar{w}) \rangle] < +\infty$  and  $\mathbb{E}[\|G_n - \nabla L(\bar{w})\|] < +\infty$ . Therefore, using assumption (A1), we obtain

$$\begin{aligned} (\forall n \in \mathbb{N}^*) \quad \mathbb{E}[\langle w_n - \bar{w}, G_n - \nabla L(\bar{w}) \rangle] &= \mathbb{E}[\mathbb{E}[\langle w_n - \bar{w}, G_n - \nabla L(\bar{w}) \rangle | \mathcal{F}_n]] \\ &= \mathbb{E}[\langle w_n - \bar{w}, \mathbb{E}[G_n - \nabla L(\bar{w}) | \mathcal{F}_n] \rangle] \\ &= \mathbb{E}[\langle w_n - \bar{w}, \nabla L(w_n) - \nabla L(\bar{w}) \rangle]. \end{aligned} \quad (4.5)$$

Moreover, using the assumption (A2), we have

$$\begin{aligned} \mathbb{E}[\|G_n - \nabla L(\bar{w})\|^2] &\leq 2\mathbb{E}[\|\nabla L(w_n) - \nabla L(\bar{w})\|^2] + 2\mathbb{E}[\|G_n - \nabla L(w_n)\|^2] \\ &\leq 2\mathbb{E}[\|\nabla L(w_n) - \nabla L(\bar{w})\|^2] + 2\sigma^2(1 + \alpha_n \mathbb{E}[\|\nabla L(w_n)\|^2]) \\ &\leq (2 + 4\sigma^2\alpha_n)\mathbb{E}[\|\nabla L(w_n) - \nabla L(\bar{w})\|^2] + 2\sigma^2(1 + 2\alpha_n\|\nabla L(\bar{w})\|^2) \\ &\leq \frac{2 + 4\sigma^2\alpha_n}{\beta} \mathbb{E}[\langle w_n - \bar{w}, \nabla L(w_n) - \nabla L(\bar{w}) \rangle] + 2\sigma^2(1 + 2\alpha_n)\|\nabla L(\bar{w})\|^2, \end{aligned} \quad (4.6)$$



where the last inequality follows from the fact that  $\nabla L$  is cocoercive since it is Lipschitz-continuous (see [4, Theorem 18.15]). We derive from (4.4), (4.5), and (4.6) that

$$\begin{aligned}\mathbb{E}[\|y_n - \bar{w}\|^2] &\leq \mathbb{E}[\|w_n - \bar{w}\|^2] - 2\gamma_n \left(1 - \frac{\gamma_n(1 + 2\sigma^2\alpha_n)}{\beta}\right) \mathbb{E}[\langle w_n - \bar{w}, \nabla L(w_n) - \nabla L(\bar{w}) \rangle] \\ &\quad + 2\sigma^2(1 + 2\alpha_n)\|\nabla L(\bar{w})\|^2 - \mathbb{E}[\|u_n\|^2].\end{aligned}\quad (4.7)$$

Therefore, combining (4.3) with (4.7), and recalling the definition of  $\varepsilon$ , we get

$$\begin{aligned}\mathbb{E}[\|w_{n+1} - \bar{w}\|^2] &\leq (1 - \lambda_n)\mathbb{E}[\|w_n - \bar{w}\|^2] + \lambda_n\mathbb{E}[\|y_n - \bar{w}\|^2] \\ &\leq \mathbb{E}[\|w_n - \bar{w}\|^2] - 2\varepsilon\gamma_n\lambda_n\mathbb{E}[\langle w_n - \bar{w}, \nabla L(w_n) - \nabla L(\bar{w}) \rangle] + 2\sigma^2\chi_n^2 - \lambda_n\mathbb{E}[\|u_n\|^2] \\ &\leq \mathbb{E}[\|w_n - \bar{w}\|^2] + 2\sigma^2\chi_n^2,\end{aligned}\quad (4.8)$$

where the last inequality follows by the monotonicity of  $\nabla L$ .

(i): Since the sequence  $(\chi_n^2)_{n \in \mathbb{N}^*}$  is summable by assumption (A4), we derive from (4.8) that  $(\mathbb{E}[\|w_{n+1} - \bar{w}\|^2])_{n \in \mathbb{N}^*}$  converges to a finite value.

(ii): We estimate the conditional expectation with respect to  $\mathcal{F}_n$  of each term in the right hand side of (4.4). Since  $w_n$  is  $\mathcal{F}_n$ -measurable, we have

$$\mathbb{E}[\|w_n - \bar{w}\|^2 | \mathcal{F}_n] = \|w_n - \bar{w}\|^2. \quad (4.9)$$

Using assumption (A1),

$$\begin{aligned}(\forall n \in \mathbb{N}^*) \quad \mathbb{E}[\langle w_n - \bar{w}, G_n - \nabla L(\bar{w}) \rangle | \mathcal{F}_n] &= \langle w_n - \bar{w}, \mathbb{E}[G_n - \nabla L(\bar{w}) | \mathcal{F}_n] \rangle \\ &= \langle w_n - \bar{w}, \nabla L(w_n) - \nabla L(\bar{w}) \rangle.\end{aligned}\quad (4.10)$$

Next, note that  $\nabla L(w_n)$  is  $\mathcal{F}_n$ -measurable by (A1), and therefore by (A2), we get

$$\begin{aligned}\mathbb{E}[\|G_n - \nabla L(\bar{w})\|^2 | \mathcal{F}_n] &\leq 2\mathbb{E}[\|\nabla L(w_n) - \nabla L(\bar{w})\|^2 | \mathcal{F}_n] + 2\mathbb{E}[\|G_n - \nabla L(w_n)\|^2 | \mathcal{F}_n] \\ &\leq 2\|\nabla L(w_n) - \nabla L(\bar{w})\|^2 + 2\sigma^2(1 + \alpha_n)\|\nabla L(w_n)\|^2 \\ &\leq 2\|\nabla L(w_n) - \nabla L(\bar{w})\|^2 + 2\sigma^2(1 + 2\alpha_n)\|\nabla L(w_n) - \nabla L(\bar{w})\|^2 + 2\alpha_n\|\nabla L(\bar{w})\|^2 \\ &\leq \frac{(2 + 4\sigma^2\alpha_n)}{\beta} \langle w_n - \bar{w}, \nabla L(w_n) - \nabla L(\bar{w}) \rangle + 2\sigma^2(1 + 2\alpha_n)\|\nabla L(\bar{w})\|^2,\end{aligned}\quad (4.11)$$

where the last inequality follows from the cocoercivity of  $\nabla L$ . Taking the conditional expectation with respect to  $\mathcal{F}_n$ , and invoking (4.4), (4.9), (4.10), and (4.11), we obtain,

$$\begin{aligned}\mathbb{E}[\|w_{n+1} - \bar{w}\|^2 | \mathcal{F}_n] &\leq (1 - \lambda_n)\|w_n - \bar{w}\|^2 + \lambda_n\mathbb{E}[\|y_n - \bar{w}\|^2 | \mathcal{F}_n] \\ &\leq \|w_n - \bar{w}\|^2 - 2\gamma_n\lambda_n \left(1 - \frac{\gamma_n(1 + 2\sigma^2\alpha_n)}{\beta}\right) \langle \nabla L(w_n) - \nabla L(\bar{w}), w_n - \bar{w} \rangle \\ &\quad + 2\sigma^2\chi_n^2 - \lambda_n\mathbb{E}[\|u_n\|^2 | \mathcal{F}_n] \\ &\leq \|w_n - \bar{w}\|^2 - 2\varepsilon\gamma_n\lambda_n \langle \nabla L(w_n) - \nabla L(\bar{w}), w_n - \bar{w} \rangle + 2\sigma^2\chi_n^2 - \lambda_n\mathbb{E}[\|u_n\|^2 | \mathcal{F}_n] \\ &\leq \|w_n - \bar{w}\|^2 + 2\sigma^2\chi_n^2.\end{aligned}\quad (4.12)$$

Hence,  $(w_n)_{n \in \mathbb{N}^*}$  is a random quasi-Fejér sequence with respect to the nonempty closed and convex set  $\text{Argmin}(L + R)$ .

Taking into account that  $\mathbb{E}[\|w_1\|^2] < +\infty$  since  $w_1 \in \mathcal{H}$  is a fixed point by assumption, it follows from Proposition A.2(i) that  $(\|w_n - \bar{w}\|^2)_{n \in \mathbb{N}^*}$  converges a.s to some integrable random variable  $\zeta_{\bar{w}}$ .

(iii): We derive from (4.8) that

$$\sum_{n \in \mathbb{N}^*} \gamma_n \lambda_n \mathbb{E}[\langle w_n - \bar{w}, \nabla L(w_n) - \nabla L(\bar{w}) \rangle] < +\infty. \quad (4.13)$$

Since  $\sum_{n \in \mathbb{N}^*} \lambda_n \gamma_n = +\infty$ , we obtain

$$\lim_{n \rightarrow \infty} \mathbb{E}[\langle w_n - \bar{w}, \nabla L(w_n) - \nabla L(\bar{w}) \rangle] = 0 \quad \Rightarrow \quad \lim_{n \rightarrow \infty} \mathbb{E}[\|\nabla L(w_n) - \nabla L(\bar{w})\|^2] = 0, \quad (4.14)$$

using again the cocoercivity of  $\nabla L$ .

(iv) We directly get from (4.8) that  $\sum_{n \in \mathbb{N}^*} \lambda_n \|u_n\|^2 < +\infty$ .

Since  $\nabla L$  is Lipschitz-continuous, and  $(\mathbb{E}[\|w_n - \bar{w}\|^2])_{n \in \mathbb{N}^*}$  is convergent by (i), there exists  $M \in ]0, +\infty[$  such that

$$(\forall n \in \mathbb{N}^*) \quad \mathbb{E}[\langle w_n - \bar{w}, \nabla L(w_n) - \nabla L(\bar{w}) \rangle] \leq \beta^{-1} \mathbb{E}[\|w_n - \bar{w}\|^2] \leq M < +\infty. \quad (4.15)$$

Hence, we derive from (4.6) and (2.3) that

$$\sum_{n \in \mathbb{N}^*} \lambda_n \gamma_n^2 \mathbb{E}[\|G_n - \nabla L(\bar{w})\|^2] < +\infty. \quad (4.16)$$

Now, recalling the definition of  $u_n$  in (4.2), using (4.16) and (4.8), we obtain

$$\sum_{n \in \mathbb{N}^*} \lambda_n \mathbb{E}[\|w_n - y_n\|^2] \leq 2 \sum_{n \in \mathbb{N}^*} \lambda_n \mathbb{E}[\|u_n\|^2] + 2 \sum_{n \in \mathbb{N}^*} \lambda_n \gamma_n^2 \mathbb{E}[\|G_n - \nabla L(\bar{w})\|^2] < +\infty. \quad (4.17)$$

□

*Proof.* [of Theorem 3.2] Since  $\mu + \nu > 0$ , then  $L + R$  is strongly convex. Hence, the problem (2.1) has a unique minimizer, i.e,  $\text{Argmin} T = \{\bar{w}\}$  for some  $\bar{w} \in \mathcal{H}$ . Moreover, since  $\gamma_n R$  is  $\gamma_n \nu$ -strongly convex, by [4, Proposition 23.11]  $\text{prox}_{\gamma_n R}$  is  $(1 + \gamma_n \nu)$ -cocoercive, and then

$$(\forall n \in \mathbb{N}^*) \quad \|y_n - \bar{w}\|^2 \leq \frac{1}{(1 + \gamma_n \nu)^2} \|(w_n - \bar{w}) - \gamma_n (G_n - \nabla L(\bar{w}))\|^2.$$

Next, proceeding as in the proof of Proposition 4.1 and recalling (4.5)-(4.6), we get an estimate analogous to the one in (4.7), namely

$$\begin{aligned} \mathbb{E}[\|y_n - \bar{w}\|^2] &\leq \frac{1}{(1 + \gamma_n \nu)^2} \left( \mathbb{E}[\|w_n - \bar{w}\|^2] - 2\gamma_n \left( 1 - \gamma_n \frac{1 + 2\sigma^2 \alpha_n}{\beta} \right) \cdot \right. \\ &\quad \left. \cdot \mathbb{E}[\langle w_n - \bar{w}, \nabla L(w_n) - \nabla L(\bar{w}) \rangle] + 2\gamma_n^2 \sigma^2 (1 + \alpha_n \|\nabla L(\bar{w})\|^2) \right). \end{aligned} \quad (4.18)$$

Since  $L$  is strongly convex of parameter  $\mu$ , it holds  $\langle \nabla L(w_n) - \nabla L(\bar{w}), w_n - \bar{w} \rangle \geq \mu \|w_n - \bar{w}\|^2$ . Therefore, from (4.18), using the  $\mu$ -strong convexity of  $L$  and recalling the definition of  $\varepsilon$ , we get

$$\mathbb{E}[\|y_n - \bar{w}\|^2] \leq \frac{1}{(1 + \gamma_n \nu)^2} \left( (1 - 2\gamma_n \mu \varepsilon) \mathbb{E}[\|w_n - \bar{w}\|^2] + 2\sigma^2 \chi_n^2 \right). \quad (4.19)$$

Hence, by definition of  $w_{n+1}$ ,

$$\mathbb{E}[\|w_{n+1} - \bar{w}\|^2] \leq \left(1 - \frac{\lambda_n \gamma_n (2\nu + \gamma_n \nu^2 + 2\mu\epsilon)}{(1 + \gamma_n \nu)^2}\right) \mathbb{E}[\|w_n - \bar{w}\|^2] + \frac{2\sigma^2 \chi_n^2}{(1 + \gamma_n \nu)^2}. \quad (4.20)$$

Let  $\gamma_n = c_1 n^{-\theta}$  and fix  $n \geq n_0$ . Since  $\gamma_n \leq \gamma_{n_0} = c_1 n_0^{-\theta} \leq 1$ , we have

$$\frac{\lambda_n \gamma_n (2\nu + \gamma_n \nu^2 + 2\mu\epsilon)}{(1 + \gamma_n \nu)^2} \geq \frac{2\lambda(\nu + \mu\epsilon)}{(1 + \nu)^2} \gamma_n = cn^{-\theta}, \quad (4.21)$$

where we set  $c = c_1 2\lambda(\nu + \mu\epsilon)/(1 + \nu)^2$ . On the other hand,

$$\frac{2\sigma^2 \chi_n^2}{(1 + \gamma_n \nu)^2} \leq 2\sigma^2 (1 + \bar{\alpha} \|\nabla L(\bar{w})\|) c_1^2 n^{-2\theta}. \quad (4.22)$$

Then, putting together (4.20), (4.21), and (4.22), we get  $\mathbb{E}[\|w_{n+1} - \bar{w}\|^2] \leq (1 - \eta_n) \mathbb{E}[\|w_n - \bar{w}\|^2] + \tau \eta_n^2$ , with  $\tau = 2\sigma^2 c_1^2 (1 + \bar{\alpha} \|\nabla L(\bar{w})\|)/c^2$  and  $\eta_n = cn^{-\theta}$ . Then, (3.1) follows from Lemma A.3.  $\square$

*Proof.* [of Theorem 3.4] Since  $L$  is uniformly convex at  $\bar{w}$ , there exists  $\phi: [0, +\infty[ \rightarrow [0, +\infty[$  increasing and vanishing only at 0 such that

$$\langle \nabla L(w_n) - \nabla L(\bar{w}), w_n - \bar{w} \rangle \geq \phi(\|w_n - \bar{w}\|). \quad (4.23)$$

Therefore, we derive from Proposition 4.1 (iii) that  $\sum_{n \in \mathbb{N}^*} \lambda_n \gamma_n \mathbb{E}[\phi(\|w_n - \bar{w}\|)] < \infty$ , and hence

$$\sum_{n \in \mathbb{N}^*} \lambda_n \gamma_n \phi(\|w_n - \bar{w}\|) < \infty \quad \text{a.s.} \quad (4.24)$$

Since  $(\lambda_n \gamma_n)_{n \in \mathbb{N}^*}$  is not summable, we have  $\liminf \phi(\|w_n - \bar{w}\|) = 0$  a.s. Consequently, there exists a subsequence  $(k_n)_{n \in \mathbb{N}^*}$  such that  $\phi(\|w_{k_n} - \bar{w}\|) \rightarrow 0$  a.s, which implies that  $\|w_{k_n} - \bar{w}\| \rightarrow 0$  a.s. In view of Proposition 4.1(ii), we get  $w_n \rightarrow \bar{w}$  a.s.  $\square$

*Proof.* [of Theorem 3.5] By Proposition 4.1(i),  $(\|w_n - \bar{w}\|^2)_{n \in \mathbb{N}^*}$  converges to an integrable random variable, hence it is uniformly bounded. Moreover,  $\liminf \mathbb{E}[\|\nabla L(w_n) - \nabla L(\bar{w})\|^2] = 0$ , and hence there exists a subsequence  $(k_n)_{n \in \mathbb{N}^*}$  such that  $\lim_{n \rightarrow \infty} \mathbb{E}[\|\nabla L(w_{k_n}) - \nabla L(\bar{w})\|^2] = 0$ . Thus, there exists a subsequence  $(p_n)_{n \in \mathbb{N}^*}$  of  $(k_n)_{n \in \mathbb{N}^*}$  such that

$$\|\nabla L(w_{p_n}) - \nabla L(\bar{w})\|^2 \rightarrow 0 \quad \text{a.s.} \quad (4.25)$$

Let  $\bar{z}$  be a weak cluster point of  $(w_{p_n})_{n \in \mathbb{N}^*}$ , then there exists a subsequence  $(w_{q_{p_n}})_{n \in \mathbb{N}^*}$  such that for almost all  $\omega$ ,  $w_{q_{p_n}}(\omega) \rightharpoonup \bar{z}(\omega)$ . Since  $\nabla L$  is weakly continuous, for almost all  $\omega$ ,  $\nabla L(w_{q_{p_n}}(\omega)) \rightharpoonup \nabla L(\bar{z}(\omega))$ . Therefore, for almost every  $\omega$ , by (4.25),  $\nabla L(\bar{w}) = \nabla L(\bar{z}(\omega))$ , and hence

$$\langle \nabla L(\bar{z}(\omega)) - \nabla L(\bar{w}), \bar{z}(\omega) - \bar{w} \rangle = 0.$$

Since  $L$  is strictly convex,  $\nabla L$  is strictly monotone, we obtain  $\bar{w} = \bar{z}(\omega)$ . This shows that  $w_{q_{p_n}} \rightharpoonup \bar{w}$  a.s.  $\square$

**Acknowledgments.** This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. L. R. acknowledges the financial support of the Italian Ministry of Education, University and Research FIRB project RBFR12M3AC. S. V. is member of the Gruppo Nazionale per l'Analisi Matematica, la Probabilit  e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM).

## REFERENCES

- [1] Y. F. Atchade, G. Fort, and E. Moulines. On stochastic proximal gradient algorithms. *arXiv:1402.2365*, February 2014.
- [2] F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Proceedings NIPS*, 2011.
- [3] Kengy Barty, Jean-Sébastien Roy, and Cyrille Strugarek. Hilbert-valued perturbed subgradient algorithms. *Math. Oper. Res.*, 32(3):551–562, 2007.
- [4] Heinz H. Bauschke and Patrick L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, New York, 2011. With a foreword by Hedy Attouch.
- [5] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.
- [6] Abdelkrim Bennar and Jean-Marie Monnez. Almost sure convergence of a stochastic approximation process in a convex set. *Int. J. Appl. Math.*, 20(5):713–722, 2007.
- [7] Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive algorithms and stochastic approximations*, volume 22 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1990. Translated from the French by Stephen S. Wilson.
- [8] Dimitri P Bertsekas. A new class of incremental gradient methods for least squares problems. *SIAM Journal on Optimization*, 7(4):913–926, 1997.
- [9] Dimitri P Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: a survey. *Optimization for Machine Learning*, page 85, 2011.
- [10] Dimitri P. Bertsekas and John N. Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM J. Optim.*, 10(3):627–642 (electronic), 2000.
- [11] Léon Bottou and Yann Le Cun. On-line learning for very large data sets. *Applied Stochastic Models in Business and Industry*, 21(2):137–151, 2005.
- [12] X. Chen and H. White. Asymptotic properties of some projection-based robbins-monro procedures in a hilbert space. *Stud. Nonlinear Dyn. Econom.*, 6:1–53, 2002.
- [13] Patrick L. Combettes. Quasi-Fejérian analysis of some optimization algorithms. In *Inherently parallel algorithms in feasibility and optimization and their applications (Haifa, 2000)*, volume 8 of *Stud. Comput. Math.*, pages 115–152. North-Holland, Amsterdam, 2001.
- [14] Patrick L. Combettes and Jean-Christophe Pesquet. Proximal thresholding algorithm for minimization over orthonormal bases. *SIAM J. Optim.*, 18(4):1351–1376, 2007.
- [15] Patrick L. Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, volume 49 of *Springer Optim. Appl.*, pages 185–212. Springer, New York, 2011.
- [16] Patrick L. Combettes and Valérie R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Model. Simul.*, 4(4):1168–1200 (electronic), 2005.
- [17] C. De Mol, E. De Vito, and L. Rosasco. Elastic-net regularization in learning theory. *J. Complexity*, 25:201–230, 2009.
- [18] O. Devolder. Stochastic first order methods in smooth convex optimization. Technical report, Center for Operations Research and econometrics, 2011.
- [19] John Duchi and Yoram Singer. Efficient online and batch learning using forward backward splitting. *J. Mach. Learn. Res.*, 10:2899–2934, 2009.
- [20] Rick Durrett. *Probability: theory and examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, fourth edition, 2010.
- [21] Yu. M. Ermol’ev and A. D. Tuniev. Random fejér and quasi-fejér sequences. *Theory of Optimal Solutions–Akademiya Nauk Ukrainskoï SSR Kiev*, 2:76–83, 1968.
- [22] Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stoch. Syst.*, 1(1):17–58, 2011.
- [23] Harold J. Kushner and Dean S. Clark. *Stochastic approximation methods for constrained and unconstrained systems*, volume 26 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1978.
- [24] J. T. Kwok, C. Hu, and W. Pan. Accelerated gradient methods for stochastic optimization and online learning. In *Advances in Neural Information Processing Systems*, volume 22, pages 781–789, 2009.
- [25] Guanghai Lan. An optimal method for stochastic composite optimization. *Math. Program.*, 133(1-2, Ser. A):365–397, 2012.
- [26] Michel Métivier. *Semimartingales*, volume 2 of *de Gruyter Studies in Mathematics*. Walter de Gruyter & Co., Berlin, 1982. A course on stochastic processes.
- [27] Jean-Marie Monnez. Almost sure convergence of stochastic gradient processes with matrix step sizes. *Statist. Probab. Lett.*, 76(5):531–536, 2006.

- [28] Sofia Mosci, Lorenzo Rosasco, Matteo Santoro, Alessandro Verri, and Silvia Villa. Solving structured sparsity regularization with proximal methods. In *ECML/PKDD (2)*, pages 418–433, 2010.
- [29] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4):1574–1609, 2008.
- [30] A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Intersci. Ser. Discrete Math. 15. John Wiley, New York, 1983.
- [31] Y. Nesterov. Gradient methods for minimizing composite objective function. CORE Discussion Paper 2007/76, Catholic University of Louvain, September 2007.
- [32] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30(4):838–855, 1992.
- [33] Alexander Rakhlin, Ohad Shamir, and Karthik Sridaran. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.
- [34] H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics (Proc. Sympos., Ohio State Univ., Columbus, Ohio, 1971)*, pages 233–257. Academic Press, New York, 1971.
- [35] Herbert Robbins and Sutton Monro. A stochastic approximation method. *Ann. Math. Statistics*, 22:400–407, 1951.
- [36] Mark W. Schmidt, Nicolas Le Roux, and Francis Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *NIPS*, pages 1458–1466, 2011.
- [37] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for svm. *Proceedings ICML*, 2007.
- [38] S. Shalev-Shwartz and N. Srebro. Svm optimization: inverse dependence on training set size. *Proceedings ICML*, 2008.
- [39] Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.
- [40] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.
- [41] Silvia Villa, Saverio Salzo, Luca Baldassarre, and Alessandro Verri. Accelerated and inexact forward-backward algorithms. *SIAM J. Optim.*, 23(3):1607–1633, 2013.
- [42] Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *J. Mach. Learn. Res.*, 11:2543–2596, 2010.
- [43] Tong Zhang. Multi-stage convex relaxation for learning with sparse regularization. In *Advances in Neural Information Processing Systems*, pages 1929–1936, 2008.
- [44] Z. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

**Appendix A. Preliminaries and auxiliary results.** The gradient of a differentiable function  $L$  from  $\mathcal{H}$  to  $\mathbb{R}$  is denoted by  $\nabla L$ . The gradient  $\nabla L$  is said to be  $\beta^{-1}$ -Lipschitz continuous, for some  $\beta^{-1} \in ]0, +\infty[$ , if

$$(\forall w \in \mathcal{H})(\forall v \in \mathcal{H}) \quad \|\nabla L(w) - \nabla L(v)\| \leq \beta^{-1} \|w - v\|. \quad (\text{A.1})$$

Let  $R: \mathcal{H} \rightarrow ]-\infty, +\infty]$  be a proper lower semicontinuous convex function, we recall that the proximity operator of  $R$  is

$$\text{prox}_R: \mathcal{H} \rightarrow \mathcal{H}: w \mapsto \underset{v \in \mathcal{H}}{\operatorname{argmin}} R(v) + \frac{1}{2} \|w - v\|^2. \quad (\text{A.2})$$

Through out this note, we assume implicitly that the closed-form expressions of the proximity operators of  $R$  are available. We refer to [4, 15] for the closed-form expression of the wide class of function in the literature. The following notions and results will be required in subsequent. We denote by  $\ell_+^1(\mathbb{N})$  the set of summable sequences in  $[0, +\infty[$ .

**DEFINITION A.1.** [21] *Let  $S$  be a non-empty subset of  $\mathcal{H}$  and let  $(\varepsilon_n)_{n \in \mathbb{N}^*}$  be sequence in  $\ell_+^1(\mathbb{N}^*)$ . Then,*

(i) A sequence  $(w_n)_{n \in \mathbb{N}^*}$  in  $\mathcal{H}$  is deterministic quasi-Fejér monotone with respect to the target set  $S$  if

$$(\forall w \in S)(\forall n \in \mathbb{N}^*) \quad \|w_{n+1} - w\|^2 \leq \|w_n - w\|^2 + \varepsilon_n. \quad (\text{A.3})$$

(ii) A sequence of random vectors  $(w_n)_{n \in \mathbb{N}^*}$  in  $\mathcal{H}$  is stochastic quasi-Fejér monotone with respect to the target set  $S$  if  $\mathbb{E}[\|w_1\|^2] < +\infty$  and

$$(\forall w \in S)(\forall n \in \mathbb{N}^*) \quad \mathbb{E}[\|w_{n+1} - w\|^2 | \sigma(w_1, \dots, w_n)] \leq \|w_n - w\|^2 + \varepsilon_n. \quad (\text{A.4})$$

PROPOSITION A.2. [3, Lemma 2.3] Let  $S$  be a non-empty closed subset of  $\mathcal{H}$ , let  $(\varepsilon_n)_{n \in \mathbb{N}^*} \in \ell_+^1(\mathbb{N}^*)$ . Let  $(w_n)_{n \in \mathbb{N}^*}$  be a sequence of random vectors in  $\mathcal{H}$  such that  $\mathbb{E}[\|w_1\|^2] < +\infty$ , and let  $\mathcal{F}_n = \sigma(w_1, \dots, w_n)$ . Assume that

$$(\forall n \in \mathbb{N}^*) \quad \mathbb{E}[\|w_{n+1} - w\|^2 | \mathcal{F}_n] \leq \|w_n - w\|^2 + \varepsilon_n \quad \text{a.s.} \quad (\text{A.5})$$

Then the following hold.

- (i) Let  $w \in S$ . Then,  $(\mathbb{E}[\|w_n - w\|^2])_{n \in \mathbb{N}^*}$  converges to some  $\zeta_w \in \mathbb{R}$  and  $(\|w_n - w\|^2)_{n \in \mathbb{N}^*}$  converges a.s. to an integrable random vector  $\xi_w$ .
- (ii)  $(w_n)_{n \in \mathbb{N}^*}$  is bounded a.s.
- (iii) The set of weak cluster points of  $(w_n)_{n \in \mathbb{N}^*}$  is non-empty a.s.
- (iv) If  $S$  is singleton, then  $(w_n)_{n \in \mathbb{N}^*}$  converges weakly a.s. to a random vector in  $S$  a.s. if and only if every its weak cluster point is in  $S$  a.s.

*Proof.* It follows from (A.5) that

$$(\forall n \in \mathbb{N}^*)(\forall w \in S) \quad \mathbb{E}[\|w_{n+1} - w\|^2] \leq \mathbb{E}[\|w_n - w\|^2] + \varepsilon_n. \quad (\text{A.6})$$

(i): Since the sequence  $(\varepsilon_n)_{n \in \mathbb{N}^*}$  is summable and  $\mathbb{E}[\|w - w_1\|^2]$  is finite, we derive from (A.6) that  $(\mathbb{E}[\|w_n - w\|^2])_{n \in \mathbb{N}^*}$  is a real positive quasi-Fejér sequence, and therefore it converges to some  $\zeta_w \in \mathbb{R}$  by [13, Lemma 3.1]. Set

$$(\forall n \in \mathbb{N}^*) \quad r_n = \|w_n - w\|^2 + \sum_{k=n}^{\infty} \varepsilon_k. \quad (\text{A.7})$$

Then, it follows from (A.5) that

$$\begin{aligned} (\forall n \in \mathbb{N}^*) \quad \mathbb{E}[r_{n+1} | \mathcal{F}_n] &= \mathbb{E}[\|w_{n+1} - w\|^2 | \mathcal{F}_n] + \sum_{k=n+1}^{\infty} \varepsilon_k \\ &\leq \|w_n - w\|^2 + \sum_{k=n}^{\infty} \varepsilon_k \\ &= r_n. \end{aligned} \quad (\text{A.8})$$

Therefore  $(r_n)_{n \in \mathbb{N}}$  is a (real) supermartingale. Since  $\sup_n \mathbb{E}[\min\{r_n, 0\}] < +\infty$  by (A.6),  $r_n$  converges a.s. to an integrable random variable [26, Theorem 9.4], that we denote by  $\xi_w$ .

(ii)&(iii): Follow directly by (i).

(iv): Clear.  $\square$

We next have the following lemma; see also [2].

LEMMA A.3. Let  $\alpha$  be in  $]0, 1]$ , and let  $c$  and  $\tau$  be in  $]0, +\infty[$ , let  $(\eta_n)_{n \in \mathbb{N}^*}$  be a strictly positive sequence defined by  $(\forall n \in \mathbb{N}^*) \eta_n = cn^{-\alpha}$ . Let  $(s_n)_{n \in \mathbb{N}^*}$  be such that

$$(\forall n \in \mathbb{N}^*) \quad 0 \leq s_{n+1} \leq (1 - \eta_n)s_n + \tau\eta_n^2. \quad (\text{A.9})$$

Let  $n_0$  be the smallest integer such that  $(\forall n \geq n_0 > 1) \eta_n \leq 1$  and set  $t = 1 - 2^{\alpha-1} \geq 0$ . Then, for every  $n \geq 2n_0$ ,

$$s_{n+1} \leq \begin{cases} \left( \tau c^2 \varphi_{1-2\alpha}(n) + s_{n_0} \exp\left(\frac{cn_0^{1-\alpha}}{1-\alpha}\right) \right) \exp\left(\frac{-ct(n+1)^{1-\alpha}}{1-\alpha}\right) + \frac{\tau 2^\alpha c}{(n-2)^\alpha} & \text{if } \alpha \in ]0, 1[, \\ s_{n_0} \left(\frac{n_0}{n+1}\right)^c + \frac{\tau c^2}{(n+1)^c} \left(1 + \frac{1}{n_0}\right)^c \varphi_{c-1}(n) & \text{if } \alpha = 1. \end{cases} \quad (\text{A.10})$$

*Proof.* Note that, for every  $m \in \mathbb{N}^*$ ,  $n \in \mathbb{N}^*$ ,  $m \leq n$ :

$$\sum_{k=m}^n k^{-\alpha} \geq \varphi_{1-\alpha}(n+1) - \varphi_{1-\alpha}(m), \quad (\text{A.11})$$

where  $\varphi_{1-\alpha}$  is defined by (1.1). Since all terms in (A.9) are positive for  $n \geq n_0$ , by applying the recursion  $n - n_0$  times we have

$$s_{n+1} \leq s_{n_0} \prod_{k=n_0}^n (1 - \eta_k) + \tau \sum_{k=n_0}^n \prod_{i=k+1}^n (1 - \eta_i) \eta_k^2. \quad (\text{A.12})$$

Let us estimate the first term in the right hand side of (A.12). Since  $1 - x \leq \exp(-x)$  for every  $x \in \mathbb{R}$ , from (A.11), we derive

$$\begin{aligned} s_{n_0} \prod_{k=n_0}^n (1 - \eta_k) &= s_{n_0} \prod_{k=n_0}^n \left(1 - \frac{c}{k^\alpha}\right) \leq s_{n_0} \exp\left(-c \sum_{k=n_0}^n k^{-\alpha}\right) \\ &\leq \begin{cases} s_{n_0} \left(\frac{n_0}{n+1}\right)^c & \text{if } \alpha = 1, \\ s_{n_0} \exp\left(\frac{c}{1-\alpha} (n_0^{1-\alpha} - (n+1)^{1-\alpha})\right) & \text{if } 0 < \alpha < 1. \end{cases} \end{aligned} \quad (\text{A.13})$$

To estimate the second term in the right hand side of (A.12), let us first consider the case  $\alpha < 1$ , let  $m \in \mathbb{N} \setminus \{0\}$  such that  $n_0 \leq n/2 \leq m+1 \leq (n+1)/2$ . We have

$$\begin{aligned} \sum_{k=n_0}^n \prod_{i=k+1}^n (1 - \eta_i) \eta_k^2 &= \sum_{k=n_0}^m \prod_{i=k+1}^n (1 - \eta_i) \eta_k^2 + \sum_{k=m+1}^n \prod_{i=k+1}^n (1 - \eta_i) \eta_k^2 \\ &\leq \exp\left(-\sum_{i=m+1}^n \eta_i\right) \sum_{k=n_0}^m \eta_k^2 + \eta_m \sum_{k=m+1}^n \left(\prod_{i=k+1}^n (1 - \eta_i) - \prod_{i=k}^n (1 - \eta_i)\right) \\ &= \exp\left(-\sum_{i=m+1}^n \eta_i\right) \sum_{k=n_0}^m \eta_k^2 + \eta_m \left(1 - \prod_{i=m+1}^n (1 - \eta_i)\right) \\ &\leq \exp\left(-\sum_{i=m+1}^n \eta_i\right) \sum_{k=n_0}^m \eta_k^2 + \eta_m \\ &\leq c^2 \exp\left(\frac{c}{1-\alpha} ((m+1)^{1-\alpha} - (n+1)^{1-\alpha})\right) \varphi_{1-2\alpha}(n) + \eta_m \quad (\text{A.14}) \\ &\leq c^2 \exp\left(\frac{-ct(n+1)^{1-\alpha}}{1-\alpha}\right) \varphi_{1-2\alpha}(n) + \frac{2^\alpha c}{\mu(n-2)^\alpha}. \quad (\text{A.15}) \end{aligned}$$

Hence, combining (A.13) and (A.15), for  $\alpha \in ]0, 1[$  we get

$$s_{n+1} \leq \left( \tau c^2 \varphi_{1-2\alpha}(n) + s_{n_0} \exp\left(\frac{cn_0^{1-\alpha}}{1-\alpha}\right) \right) \exp\left(\frac{-ct(n+1)^{1-\alpha}}{1-\alpha}\right) + \frac{\tau 2^{\alpha} c}{(n-2)^{\alpha}}. \quad (\text{A.16})$$

We next estimate the second term in the right hand side of (A.12) in the case  $\alpha = 1$ . We have

$$\sum_{k=n_0}^n \prod_{i=k+1}^n (1 - \eta_i) \eta_k^2 = \frac{c^2}{(n+1)^c} \left(1 + \frac{1}{n_0}\right)^c \sum_{k=n_0}^n \frac{1}{k^{2-c}} \leq \frac{c^2}{(n+1)^c} \left(1 + \frac{1}{n_0}\right)^c \varphi_{c-1}(n). \quad (\text{A.17})$$

Therefore, for  $\alpha = 1$ , we obtain,

$$s_{n+1} \leq s_{n_0} \left(\frac{n_0}{n+1}\right)^c + \frac{\tau c^2}{(n+1)^c} \left(1 + \frac{1}{n_0}\right)^c \varphi_{c-1}(n), \quad (\text{A.18})$$

which completes the proof.  $\square$

LEMMA A.4. [16, Lemma 2.4] Let  $R \in \Gamma_0(\mathcal{H})$ . Then the proximity of  $R$  is firmly non-expansive, i.e.,

$$(\forall w \in \mathcal{H})(u \in \mathcal{H}) \quad \|\text{prox}_R w - \text{prox}_R u\|^2 \leq \|u - w\|^2 - \|(w - \text{prox}_R w) - (u - \text{prox}_R u)\|^2. \quad (\text{A.19})$$

LEMMA A.5 (**Baillon-Haddad theorem**). Let  $L \in \Gamma_0(\mathcal{H})$  be a convex differentiable function with  $\beta^{-1}$  Lipschitzian gradient. Then,  $\nabla L$  is  $\beta$ -cocoercive, i.e.,

$$(\forall u \in \mathcal{H})(w \in \mathcal{H}) \quad \langle w - u, \nabla L(w) - \nabla L(u) \rangle \geq \beta \|\nabla L(w) - \nabla L(u)\|^2. \quad (\text{A.20})$$

*Proof.* [of Remark 1] Let  $w$  be a weak cluster point of  $(w_n)_{n \in \mathbb{N}^*}$ , i.e., there exists a subsequence  $(w_{k_n})_{n \in \mathbb{N}^*}$  such that  $w_{k_n} \rightharpoonup w$  a.s. Since  $T = L + R$  is convex and lower semicontinuous, it is weakly lower semicontinuous, hence

$$T(w) \leq \underline{\lim} T(w_{k_n}) = \inf T, \quad (\text{A.21})$$

which shows that  $w \in \text{Argmin } T$  a.s. In view of Proposition A.2(iv), we conclude that  $(w_n)_{n \in \mathbb{N}^*}$  converges weakly to optimal solution a.s.  $\square$

## Appendix B. Minimization over orthonormal bases.

We next describe how to apply our algorithm to minimization over orthonormal bases. This problem often arises in sparse signal recovery as well as learning theory (see e.g. [14, 40]).

PROBLEM B.1. Let  $\beta$  be in  $]0, +\infty[$ , let  $\nu$  be in  $[0, +\infty]$ , let  $(e_k)_{k \in \mathbb{N}}$  be an orthonormal base of  $\mathcal{H}$ , let  $(\phi_k)_{k \in \mathbb{N}}$  be a sequence of functions in  $\Gamma_0(\mathbb{R}_+)$  such that  $(\forall k \in \mathbb{N}) \phi_k \geq \phi_k(0) = 0$ . Let  $L$  be in  $\Gamma_0(\mathcal{H})$  such that  $L$  is differentiable with a  $\beta^{-1}$ -Lipschitz continuous gradient. The problem is to

$$\underset{w \in \mathcal{H}}{\text{minimize}} L(w) + \sum_{k \in \mathbb{N}} \left( \phi_k(\langle w, e_k \rangle) + \frac{\nu}{2} |\langle w, e_k \rangle|^2 \right). \quad (\text{B.1})$$



In the case when  $\nu = 0$  and  $L$  is non-strongly convex, we assume that the set of solutions to (B.1) is non-empty

COROLLARY B.2. Fix  $w_1 \in \mathcal{H}$  and set

$$(\forall n \in \mathbb{N}^*) \quad \begin{cases} \text{For } k = 0, 1, \dots, \\ z_{n,k} = w_{n,k} - \gamma_n \langle G_n, e_k \rangle \\ y_{n,k} = \text{prox}_{\frac{\gamma_n}{1+\nu\gamma_n}\phi_k}((1+\nu\gamma_n)^{-1}z_{n,k}) \\ w_{n+1,k} = (1-\lambda_n)w_{n,k} + \lambda_n y_{n,k}, \end{cases} \quad (\text{B.2})$$

where we assume that conditions (A1), (A2), and (A3) are satisfied. Then the following hold for some solution  $\bar{w}$  to (B.1).

- (i) If  $\mu + \nu > 0$ , then  $\bar{w}$  is unique and (3.1) holds for  $\inf_{n \in \mathbb{N}^*} \lambda_n \geq \bar{\lambda} > 0$ ,  $\sup_{n \in \mathbb{N}^*} \alpha_n \leq \bar{\alpha} < +\infty$  and  $\gamma_n = c_1 n^{-\theta}$  for some  $\theta \in ]0, 1]$  and for some  $c_1 \in ]0, +\infty[$  such that  $c = 2c_1\lambda(\nu + \mu\varepsilon)/(1+\nu)^2 \geq 1$ , and for  $\tau = 2\sigma^2 c_1^2(1 + \bar{\alpha}\|\nabla L(\bar{w})\|)/c^2$ .
- (ii) If  $L$  is uniformly convex and (A4) is also satisfied, then  $w_n \rightarrow \bar{w}$  a.s.
- (iii) If  $L$  is strictly convex and  $\nabla L$  weakly continuous, and (A4) is also satisfied then there exists a subsequence  $(t_n)_{n \in \mathbb{N}}$  such that  $w_{t_n} \rightarrow \bar{w}$  a.s.
- (iv) Define  $R: w \mapsto \sum_{k \in \mathbb{N}} (\phi_k(\langle w, e_k \rangle) + \frac{\nu}{2} |\langle w, e_k \rangle|^2)$ . Suppose that  $((L+R)(w_n))_{n \in \mathbb{N}}$  converges a.s to the optimal value noted by  $\bar{L} + \bar{R}$ , (A4) is also satisfied then the whole sequence  $(w_n)_{n \in \mathbb{N}}$  converges weakly a.s to  $\bar{w}$ .

*Proof.* Set  $R: w \mapsto \sum_{k \in \mathbb{N}^*} \phi_k(\langle w, e_k \rangle) + \frac{\nu}{2} |\langle w, e_k \rangle|^2$  is  $\nu$ -strongly convex by Parserval's identity, and its proximity operator is computable [4]. More precisely, it follows from [4, Proposition 23.29(i) and Proposition 23.34] that the iteration (2.2) reduces to (B.2). Therefore, the statement follows from Theorem 3.2, and 3.4.  $\square$